09/934,799 #10

The Patent Office

INVESTOR IN PEOPLE

CERTIFIED COPY OF PRIORITY DOCUMENT

The Patent Office
Concept House
Cardiff Road
Newport
South Wales
NP10 8QQ

I, the undersigned, being an officer duly authorised in accordance with Section 74(1) and (4) of the Deregulation & Contracting Out Act 1994, to sign and issue certificates on behalf of the Comptroller-General, hereby certify that annexed hereto is a true copy of the documents as originally filed in connection with the patent application identified therein.

accordance with the Patents (Companies Re-registration) Rules 1982, if a company named this certificate and any accompanying documents has re-registered under the Companies Act 80 with the same name as that with which it was registered immediately before re-istration save for the substitution as, or inclusion as, the last part of the name of the words ıblic limited company" or their equivalents in Welsh, references to the name of the company his certificate and any accompanying documents shall be treated as references to the name ı which it is so re-registered.

cordance with the rules, the words "public limited company" may be replaced by p.l.c., P.L.C. or PLC.

gistration under the Companies Act does not constitute a new legal entity but merely ts the company to certain additional company law rules.

Signed

Dated 9 April 2003

Patents Form 7/77

Patents Act 1977

(Rule 15)

The Patent Office

Statement of inventorship and of right to grant of a patent

The Patent Office
Cardiff Road
Newport
Gwent NP9 1RH

1. Your reference
   2644001/AM

9905186.4

:tion Number
'application reference 2644001

5 MAR 1999

3. Full name of the or each applicant

   Canon Kabushiki Kaisha

4. Title of the invention
   DATABASE ANNOTATION AND RETRIEVAL

5. State how the applicant(s) derived the right from the inventor(s) to be granted a patent

   By employment of the inventors by Canon Research Centre Europe Limited, and by a general agreement dated 1 January 1994 between Canon Research Centre Europe Limited and the applicant.

6. How many, if any additional Patents Forms 7/77 are attached to this form?

   NONE

11. I/We believe that the person(s) named over the page (and on any extra copies of this form) is/are the inventor(s) of the invention which the above patent application relates to.

Signature Beresford & Co Date 5 March 1999
BERESFORD & Co

12. Name and daytime telephone number of person to contact in the United Kingdom

ALAN MACDOUGALL

Tel: 0171-831-2290

Patents Form 7/77

CHARLESWORTH, Jason Peter Andrew
c/o Canon Research Centre Europe
Limited
1 Occam Court, Occam Road
Surrey Research Park
Guildford
Surrey GU2 5YJ 761570180 1

RAJAN, Jebu Jacob
c/o Canon Research Centre Europe
Limited
1 Occam Court, Occam Road
Surrey Research Park
Guildford
Surrey GU2 5YJ 761571900 1

GARNER, Philip Neil
c/o Canon Research Centre Europe
Limited
1 Occam Court, Occam Road
Surrey Research Park
Guildford
Surrey GU2 5YJ 761572700 1

08MAR99 E430751-1 D02917
P01/7700 0.00 - 9905186.4

Patents Form 1/77
Patents Act 1977
(Rule 16)

THE PATENT OFFICE
D
-5 MAR 1999
LONDON

The Patent Office

Request for grant of a patent

The Patent Office
Cardiff Road
Newport
Gwent NP9 1RH

1. Your reference
2644001/AM

2. Patent Application Number
9905186.4
5 MAR 1999

3. Full name, address and postcode of the or of each applicant *(underline all surnames)*

Canon Kabushiki Kaisha
30-2 3-Chome Shimomaruko
Ohta-Ku
Tokyo
Japan

Patents ADP number *(if known)* 363010003-

If the applicant is a corporate body, give the country/state of its incorporation
Country: JAPAN
State:

4. Title of the invention
DATABASE ANNOTATION AND RETRIEVAL

5. Name of agent **Beresford & Co**

"Address for Service" in the United Kingdom to which all correspondence should be sent

2/5 Warwick Court
High Holborn
London WC1R 5DJ

Patents ADP number 1826001

6. Priority details

Country	Priority application number	Date of filing

Patents Form 1/77

7. If this application is divided or otherwise derived from an earlier UK application give details

Number of earlier of application   Date of filing

8. Is a statement of inventorship and or right to grant of a patent required in support of this request?

YES

9. Enter the number of sheets for any of the following items you are filing with this form.

Continuation sheets of this form	
Description	27
Claim(s)	8
Abstract	1
Drawing(s)	14 ✗ 14

8

10. If you are also filing any of the following, state how many against each item.

Priority documents	
Translations of priority documents	
Statement of inventorship and right to grant of a patent *(Patents form 7/77)*	1 + 3 COPIES
Request for preliminary examination and search *(Patents Form 9/77)*	
Request for Substantive Examination *(Patents Form 10/77)*	
Any other documents *(please specify)*	

11. I/We request the grant of a patent on the basis of this application

Signature Beresford & Co   Date 5 March 1999
BERESFORD & Co

12. Name and daytime telephone number of person to contact in the United Kingdom   ALAN MACDOUGALL

Tel:0171-831-2290

DATABASE ANNOTATION AND RETRIEVAL

The present invention relates to the annotation of data files which are to be stored in a database for facilitating their subsequent retrieval. The present invention is also concerned with a system for generating the annotation data which is added to the data file and to a system for searching the annotation data in the database to retrieve a desired data file in response to a user's input query.

Databases of information are well known and suffer from the problem of how to locate and retrieve the desired information from the database quickly and efficiently. Existing database search tools allow the user to search the database using typed keywords. Whilst this is quick and efficient, this type of searching is not suitable for various kinds of databases, such as video or audio databases.

According to one aspect, the present invention aims to provide a data structure which will allow the annotation of data files within a database which will allow a quick and efficient search to be carried out in response to a user's input query.

Exemplary embodiments of the present invention will now

be described with reference to Figures 1 to 10, in which:

Figure 1 is a schematic view of a computer which is programmed to operate an embodiment of the present invention;

Figure 2 is a block diagram showing a phoneme and word annotator unit which is operable to generate phoneme and word annotation data for appendage to a data file;

Figure 3 is a block diagram illustrating one way in which the phoneme and word annotator can generate the annotation data from an input video data file;

Figure 4a is a schematic diagram of a phoneme lattice for an example audio string from the input video data file;

Figure 4b is a schematic diagram of a word and phoneme lattice embodying one aspect of the present invention, for an example audio string from the input video data file;

Figure 5 is a schematic block diagram of a user's terminal which allows the user to retrieve information from the database by a voice query;

Figure 6a is a flow diagram illustrating part of the flow

control of the user terminal shown in Figure 5;

Figure 6b is a flow diagram illustrating the remaining part of the flow control of the user terminal shown in Figure 5;

Figure 7 is a flow diagram illustrating the way in which a search engine forming part of the user's terminal carries out a phoneme search within the database;

Figure 8 is a schematic diagram illustrating the form of a phoneme string and four M-GRAMS generated from the phoneme string;

Figure 9 is a plot showing two vectors and the angle between the two vectors; and

Figure 10 is a schematic diagram of a pair of word and phoneme lattices, for example audio strings from two speakers;

Figure 11 is a schematic block diagram illustrating a user terminal which is operable to access a database located on a remote server via a data network in response to an input utterance by the user;

Figure 12 is a schematic block diagram of a user terminal

which allows a user to access a database located in a remote server in response to an input utterance from the user;

Figure 13 is a schematic block diagram of a user terminal which allows a user to access a database by a typed input query; and

Figure 14 is a schematic block diagram illustrating the way in which a phoneme and word lattice can be generated from script data contained within a video data file.

Embodiments of the present invention can be implemented using dedicated hardware circuits, but the embodiment to be described is implemented in computer software or code, which is run in conjunction with processing hardware such as a personal computer, work station, photocopier, facsimile machine, personal digital assistant (PDA) or the like.

Figure 1 shows a personal computer (PC) 1 which is programmed to operate an embodiment of the present invention. A keyboard 3, a pointing device 5, a microphone 7 and a telephone line 9 are connected to the PC 1 via an interface 11. The keyboard 3 and pointing device 5 enable the system to be controlled by a user. The microphone 7 converts acoustic speech signals from

the user into equivalent electrical signals and supplies them to the PC 1 for processing. An internal modem and speech receiving circuit (not shown) is connected to the telephone line 9 so that the PC 1 can communicate with, for example, a remote computer or with a remote user.

The programme instructions which make the PC 1 operate in accordance with the present invention may be supplied for use with an existing PC 1 on, for example, a storage device such as a magnetic disc 13, or by downloading the software from the Internet (not shown) via the internal modem and telephone line 9.

DATA FILE ANNOTATION

Figure 2 is a block diagram illustrating the way in which annotation data 21 for an input data file 23 is generated by a phoneme and word annotating unit 25. As shown, the generated phoneme and word annotation data 21 is then combined with the data file 23 in the data combination unit 27 and the combined data file output thereby is input to the database 29. In this embodiment, the annotation data 21 comprises a combined phoneme (or phoneme like) and word lattice which allows the user to retrieve information from the database by a voice query. As those skilled in the art will appreciate, the data file 23 can be any kind of data file, such as, a video file, an audio file, a multimedia file etc.

A system has been proposed to generate N-Best word lists for an audio stream as annotation data by passing the audio data from a video data file through an automatic speech recognition unit. However, such word-based systems suffer from a number of problems. These include (i) that state of the art speech recognition systems still make basic mistakes in recognition; (ii) that state of the art automatic speech recognition systems use a dictionary of perhaps 20,000 to 100,000 words and cannot produce words outside that vocabulary; and (iii) that the production of N-Best lists grows exponentially with the number of hypothesis at each stage, therefore resulting in the annotation data becoming prohibitively large for long utterances.

Whilst the first of these problems may not be that significant if the same automatic speech recognition system is used to generate the annotation data and to subsequently retrieve the corresponding data file, since the same decoding error could occur. However, with advances in automatic speech recognition systems being made each year, it is likely that in the future the same type of error may not occur, resulting in the inability to be able to retrieve the corresponding data file at that later date. With regard to the second problem, this is particularly significant in video data applications, since users are likely to use names and places (which may

not be in the speech recognition dictionary) as input query terms. In place of these names, the automatic speech recognition system will typically replace the out of vocabulary words with a phonetically similar word or words within the vocabulary, often corrupting nearby decodings. This can also result in the failure to retrieve the required data file upon subsequent request.

In contrast, with the proposed phoneme and word lattice annotation data, a quick and efficient search using the word data in the database 29 can be carried out and, if this fails to provide the required data file, then a further search using the more robust phoneme data can be performed. The phoneme and word lattice is an acyclic directed graph with a single entry point and a single exit point. It represents different parses of the audio stream within the data file. It is not simply a sequence of words with alternatives since each word does not have to be replaced by a single alternative, one word can be substituted for two or more words or phonemes, and the whole structure can form a substitution for one or more words or phonemes. Therefore, the density of data within the phoneme and word lattice essentially remains linear throughout the audio data, rather than growing exponentially as in the case of the N-Best technique discussed above. As those skilled in the art of speech recognition will realise, the use of phoneme data is more

robust, because phonemes are dictionary independent and allow the system to cope with out of vocabulary words, such as names, places, foreign words etc. The use of phoneme data is also capable of making the system future proof, since it allows data files which are placed into the database to be retrieved even when the words were not understood by the original automatic speech recognition system.

The way in which this phoneme and word lattice annotation data can be generated for a video data file will now be described with reference to Figure 3. As shown, the video data file 31 comprises video data 31-1, which defines the sequence of images forming the video sequence and audio data 31-2, which defines the audio which is associated with the video sequence. As is well known, the audio data 31-2 is time synchronised with the video data 31-1 so that, in use, both the video and audio data are supplied to the user at the same time.

As shown in Figure 3, in this embodiment, the audio data 31-2 is input to an automatic speech recognition unit 33, which is operable to generate a phoneme lattice corresponding to the stream of audio data 31-2. Such an automatic speech recognition unit 33 is commonly available in the art and will not be described in further detail. The reader is referred to, for example, the book

entitled 'Fundamentals of Speech Recognition' by Lawrence Rabiner and Biing-Hwang Juang and, in particular, to pages 42 to 50 thereof, for further information on this type of speech recognition system.

Figure 4a illustrates the form of the phoneme lattice data output by the speech recognition unit 33, for the input audio corresponding to the phrase '...tell me about Jason...'. As shown, the automatic speech recognition unit 33 identifies a number of different possible phoneme strings which correspond to this input audio utterance. For example, the speech recognition system considers that the first phoneme in the audio string is either a /t/ or a /d/. As is well known in the art of speech recognition, these different possibilities can have their own weighting which is generated by the speech recognition unit 33 and is indicative of the confidence of the speech recognition unit's output. For example, the phoneme /t/ may be given a weighting of 0.9 and the phoneme /d/ may be given a weighting of 0.1, indicating that the speech recognition system is fairly confident that the corresponding portion of audio represents the phoneme /t/, but that it still may be the phoneme /d/. In this embodiment, however, this weighting of the phonemes is not performed.

As shown in Figure 3, the phoneme lattice data 35 output

by the automatic speech recognition unit 33 is input to a word decoder 37 which is operable to identify possible words within the phoneme lattice data 35. In this embodiment, the words identified by the word decoder 37 are incorporated into the phoneme lattice data structure. For example, for the phoneme lattice shown in Figure 4a, the word decoder 37 identifies the words 'tell', 'dell', 'term', 'me', 'a', 'boat', 'about', 'chase' and 'sun'. As shown in Figure 4b, these identified words are added to the phoneme lattice data structure output by the speech recognition unit 33, to generate a phoneme and word lattice data structure which forms the annotation data 31-3. This annotation data 31-3 is then combined with the video data file 31 to generate an augmented video data file 31' which is then stored in the database 29. As those skilled in the art will appreciate, in a similar way to the way in which the audio data 31-2 is time synchronised with the video data 31-1, the annotation data 31-3 is also time synchronised and associated with the corresponding video data 31-1 and audio data 31-2, so that a desired portion of the video and audio data can be retrieved by searching for and locating the corresponding portion of the annotation data 31-3.

In this embodiment, the annotation data 31-3 stored in the database 29 has the following general form:

HEADER

- time of start
- flag if word if phoneme if mixed
- time index associating the location of blocks of annotation data within memory to a given time point.
- word set used (i.e. the dictionary)
- phoneme set used
- the language to which the vocabulary pertains

Block(i) i = 0,1,2,.....

node N_j j = 0,1,2,.....

- time offset of node from start of block
- phoneme links (k) k = 0,1,2.....
  offset to node $N_j = N_k - N_j$ (N_k is node to which link K extends)
  phoneme associated with link (k)
- word links (l) l = 0,1,2,.....
  offset to node $N_j = N_i - N_j$ (N_j is node to which link l extends)
  word associated with link (l)

The time of start data in the header can identify the time and date of transmission of the data. For example, if the video file is a news broadcast, then the time of start may include the exact time of the broadcast and the

date on which it was broadcast.

The flag identifying if the annotation data is word annotation data, phoneme annotation data or if it is mixed is provided since not all the data files within the database will include the combined phoneme and word lattice annotation data discussed above, and in this case, a different search strategy would be used to search this annotation data.

In this embodiment, the annotation data is divided into blocks in order to allow the search to jump into the middle of the annotation data for a given audio data stream. The header therefore includes a time index which associates the location of the blocks of annotation data within the memory to a given time offset between the time of start and the time corresponding to the beginning of the block.

The header also includes data defining the word set used (i.e. the dictionary), the phoneme set used and the language to which the vocabulary pertains. The header may also include details of the automatic speech recognition system used to generate the annotation data and any appropriate settings thereof which were used during the generation of the annotation data.

The blocks of annotation data then follow the header and identify, for each node in the block, the time offset of the node from the start of the block, the phoneme links which connect that node to other nodes by phonemes and word links which connect that node to other nodes by words. Each phoneme link and word link identifies the phoneme or word which is associated with the link. They also identify the offset to the current node. For example, if node N_{50} is linked to node N_{55} by a phoneme link, then the offset to node N_{50} is 5. As those skilled in the art will appreciate, using an offset indication like this allows the division of the continuous annotation data into separate blocks.

In an embodiment where an automatic speech recognition unit outputs weightings indicative of the confidence of the speech recognition units output, these weightings or confidence scores would also be included within the data structure. In particular, a confidence score would be provided for each node which is indicative of the confidence of arriving at the node and each of the phoneme and word links would include a transition score depending upon the weighting given to the corresponding phoneme or word. These weightings would then be used to control the search and retrieval of the data files by discarding those matches which have a low confidence score.

DATA FILE RETRIEVAL

Figure 5 is a block diagram illustrating the form of a user terminal 59 which can be used to retrieve the annotated data files from the database 29. This user terminal 59 may be, for example, a personal computer, hand held device or the like. As shown, in this embodiment, the user terminal 59 comprises the database 29 of annotated data files, an automatic speech recognition unit 51, a search engine 53, a control unit 55 and a display 57. In operation, the automatic speech recognition unit 51 is operable to process an input voice query from the user 39 received via the microphone 7 and the input line 61 and to generate therefrom corresponding phoneme and word data. This data may also take the form of a phoneme and word lattice, but this is not essential. This phoneme and word data is then input to the control unit 55 which is operable to initiate an appropriate search of the database 29 using the search engine 53. The results of the search, generated by the search engine 53, are then transmitted back to the control unit 55 which analyses the search results and generates and displays appropriate display data to the user via the display 57.

Figures 6a and 6b are flow diagrams which illustrate the way in which the user terminal 59 operates in this embodiment. In step s1, the user terminal 59 is in an

idle state and awaits an input query from the user 39. Upon receipt of an input query, the phoneme and word data for the input query is generated in step s3 by the automatic speech recognition unit 51. The control unit 55 then instructs the search engine 53, in step s5, to perform a search in the database 29 using the word data generated for the input query. The word search employed in this embodiment is the same as is currently being used in the art for typed keyword searches, and will not be described in more detail here. If in step s7, the control unit 55 identifies from the search results, that a match for the user's input query has been found, then it outputs the search results to the user via the display 57.

In this embodiment, the user terminal 59 then allows the user to consider the search results and awaits the user's confirmation as to whether or not the results correspond to the information the user requires. If they are, then the processing proceeds from step s11 to the end of the processing and the user terminal 59 returns to its idle state and awaits the next input query. If, however, the user indicates (by, for example, inputting an appropriate voice command) that the search results do not correspond to the desired information, then the processing proceeds from step s11 to step s13, where the search engine 53 performs a phoneme search of the database 29. However,

in this embodiment, the phoneme search performed in step s13 is not of the whole database 29, since this could take several hours depending on the size of the database 29.

Instead, the phoneme search performed in step s13 uses the results of the word search performed in step s5 to identify one or more portions within the database which may correspond to the user's input query. The way in which the phoneme search performed in step s13 is performed in this embodiment, will be described in more detail later. After the phoneme search has been performed, the control unit 55 identifies, in step s15, if a match has been found. If a match has been found, then the processing proceeds to step s17 where the control unit 55 causes the search results to be displayed to the user on the display 57. Again, the system then awaits the user's confirmation as to whether or not the search results correspond to the desired information. If the results are correct, then the processing passes from step s19 to the end and the user terminal 59 returns to its idle state and awaits the next input query. If however, the user indicates that the search results do not correspond to the desired information, then the processing proceeds from step s19 to step s21, where the control unit 55 is operable to ask the user, via the display 57, whether or not a phoneme search should be

performed of the whole database 29. If in response to this query, the user indicates that such a search should be performed, then the processing proceeds to step s23 where the search engine performs a phoneme search of the entire database 29.

On completion of this search, the control unit 55 identifies, in step s25, whether or not a match for the user's input query has been found. If a match is found, then the processing proceeds to step s27 where the control unit 55 causes the search results to be displayed to the user on the display 57. If the search results are correct, then the processing proceeds from step s29 to the end of the processing and the user terminal 59 returns to its idle state and awaits the next input query. If, on the other hand, the user indicates that the search results still do not correspond to the desired information, then the processing passes to step s31 where the control unit 55 queries the user, via the display 57, whether or not the user wishes to redefine or amend the search query. If the user does wish to redefine or amend the search query, then the processing returns to step s3 where the user's subsequent input query is processed in a similar manner. If the search is not to be redefined or amended, then the search results and the user's initial input query are discarded and the user terminal 59 returns to its idle state and awaits the next input

query.

PHONEME SEARCH

As mentioned above, in steps s13 and s23, the search engine 53 compares the phoneme data of the input query with the phoneme data in the phoneme and word lattice annotation data stored in the database 29. Various techniques can be used including standard pattern matching techniques such as dynamic programming, to carry out this comparison. In this embodiment, a technique which we refer to as M-GRAMS is used. This technique was proposed by Ng, K. and Zue, V.W. and is discussed in, for example, the paper entitled "Subword unit representations for spoken document retrieval" published in the proceedings of Eurospeech 1997.

The problem with searching for individual phonemes is that there will be many occurrences of each phoneme within the database. Therefore, an individual phoneme on its own does not provide enough discriminability to be able to match the phoneme string of the input query with the phoneme strings within the database. Syllable sized units, however, are likely to provide more discriminability, although they are not easy to identify. The M-GRAM technique presents a suitable compromise between these two possibilities and takes overlapping fixed size fragments, or M-GRAMS, of the phoneme string

to provide a set of features. This is illustrated in Figure 8, which shows part of an input phoneme string having phonemes a, b, c, d, e, and f, which are split into four M-GRAMS (a, b, c), (b, c, d), (c, d, e) and (d, e, f). In this illustration, each of the four M-GRAMS comprises a sequence of three phonemes which is unique and represents a unique feature (f_i) which can be found within the input phoneme string.

Therefore, referring to Figure 7, the first step s51 in performing the phoneme search in step s13 shown in Figure 6, is to identify all the different M-GRAMS which are in the input phoneme data and their frequency of occurrence. Then, in step s53, the search engine 53 determines the frequency of occurrence of the identified M-GRAMS in the selected portion of the database (identified from the word search performed in step s5 in Figure 6). To illustrate this, for a given portion of the database and for the example M-GRAMS illustrated in Figure 8, this yields the following table of information:

M-GRAM (feature (f_i))	Input phoneme string frequency of occurrence (<u>q</u>)	Phoneme string of selected portion of database (<u>a</u>)
M_1	1	0
M_2	2	2

M_3	3	2
M_4	1	1

Next, in step s55, the search engine 53 calculates a similarity score representing a similarity between the phoneme string of the input query and the phoneme string of the selected portion from the database. In this embodiment, this similarity score is determined using a cosine measure using the frequencies of occurrence of the identified M-GRAMS in the input query and in the selected portion of the database as vectors. The philosophy behind this technique is that if the input phoneme string is similar to the selected portion of the database phoneme string, then the frequency of occurrence of the M-GRAM features will be similar for the two phoneme strings. Therefore, if the frequencies of occurrence of the M-GRAMS are considered to be vectors (i.e. considering the second and third columns in the above table as vectors), then if there is a similarity between the input phoneme string and the selected portion of the database, then the angle between these vectors should be small. This is illustrated in Figure 9 for two-dimensional vectors $\underline{a}$ and $\underline{q}$, with the angle between the vectors given as θ. In the example shown in Figure 8, the vectors $\underline{a}$ and $\underline{q}$ will be four dimensional vectors and the similarity score can be calculated from:

$$SCORE = COS\ \theta = \frac{\underline{a} \cdot \underline{q}}{|\underline{a}|\ |\underline{q}|}$$

This score is then associated with the current selected portion of the database and stored until the end of the search. In some applications, the vectors used in the calculation of the cosine measure will be the logarithm of these frequencies of occurrences, rather than the frequencies of occurrences themselves.

The processing then proceeds to step s57 where the search engine 53 identifies whether or not there are any more selected portions of phoneme strings from the database 29. If there are, then the processing returns to step s53 where a similar procedure is followed to identify the score for this portion of the database. If there are no more selected portions, then the searching ends and the processing returns to step s15 shown in Figure 6, where the control unit considers the scores generated by the search engine 53 and identifies whether or not there is a match by, for example, comparing the calculated scores with a predetermined threshold value.

As those skilled in the art will appreciate, a similar matching operation will be performed in step s23 shown in Figure 6. However, since the entire database is being

searched, this search is carried out by searching each of the blocks discussed above in turn.

ALTERNATIVE EMBODIMENTS

As those skilled in the art will appreciate, this type of phonetic and word annotation of data files in a database provides a convenient and powerful way to allow a user to search the database by voice. In the illustrated embodiment, a single audio data stream was annotated and stored in the database for subsequent retrieval by the user. As those skilled in the art will appreciate, when the input data file corresponds to a video data file, the audio data within the data file will usually include audio data for different speakers. Instead of generating a single stream of annotation data for the audio data, separate phoneme and word lattice annotation data can be generated for the audio data of each speaker. This may be achieved by identifying, from the pitch or from another distinguishing feature of the speech signals, the audio data which corresponds to each of the speakers and then by annotating the different speaker's audio separately. This may also be achieved if the audio data was recorded in stereo or if an array of microphones were used in generating the audio data, since it is then possible to process the audio data to extract the data for each speaker.

Figure 10 illustrates the form of the annotation data in such an embodiment, where a first speaker utters the words "... this so" and the second speaker replies "yes". As illustrated, the annotation data for the different speakers' audio data are time synchronised, relative to each other, so that the annotation data is still time synchronised to the video and audio data within the data file. In such an embodiment, the header information in the data structure should preferably include a list of the different speakers within the annotation data and, for each speaker, data defining that speaker's language, accent, dialect and phonetic set, and each block should identify those speakers that are active in the block.

In the above embodiments, a speech recognition system was used to generate the annotation data for annotating a data file in the database. As those skilled in the art will appreciate, other techniques can be used to generate this annotation data. For example, a human operator can listen to the audio data and generate a phonetic and word transcription to thereby manually generate the annotation data.

In the above embodiment, the database 29 and the automatic speech recognition unit were both located within the user terminal 59. As those skilled in the art will appreciate, this is not essential. Figure 11

illustrates an embodiment in which the database 29 and the search engine 53 are located in a remote server 60 and in which the user terminal 59 accesses the database 29 via the network interface units 67 and 69 and a data network 68 (such as the internet). In operation, the user inputs a voice query via the microphone 7 which is converted into phoneme and word data by the automatic speech recognition unit 51. This data is then passed to the control unit which controls the transmission of this phoneme and word data over the data network 68 to the search engine 53 located within the remote server 60. The search engine 53 then carries out the search in a similar manner to the way in which the search was performed in the first embodiment. The results of the search are then transmitted back from the search engine 53 to the control unit 55 via the data network 68. The control unit considers the search results received back from the network and displays appropriate data on the display 57 for viewing by the user 39.

In addition to locating the database 29 and the search engine 53 in the remote server 60, it is also possible to locate the automatic speech recognition unit 51 in the remote server 60. Such an embodiment is shown in Figure 12. As shown in this embodiment, the input voice query from the user is passed via input line 61 to a speech encoding unit 73 which is operable to encode the speech

for efficient transfer through the data network 68. The encoded data is then passed to the control unit 55 which transmits the data over the network 68 to the remote server 60, where it is processed by the automatic speech recognition unit 51. The phoneme and word data generated by the speech recognition unit 51 for the input query is then passed to the search engine 53 for use in searching the database 29. The search results generated by the search engine 53 are then passed, via the network interface 69 and the network 68, back to the user terminal 59. The search results received back from the remote server are passed via the network interface unit 67 to the control unit 55 which analyses the search results and generates and displays appropriate data on the display 57 for viewing by the user.

In the above embodiments, the user inputs his query by voice. Figure 13 shows an alternative embodiment in which the user inputs the query via the keyboard 3. As shown, the text input via the keyboard 3 is passed to phonetic transcription unit 75 which is operable to generate a corresponding phoneme string from the input text. This phoneme string together with the words input via the keyboard 3 are then passed to the control unit 55 which initiates a search of database using the search engine 53. The way in which this search is carried out is the same as in the first embodiment and will not,

therefore, be described again. As with the other embodiments discussed above, the phonetic transcription unit 75, search engine 53 and/or the database 29 may all be located in a remote server.

In the first embodiment, the audio data from the data file 31 was passed through an automatic speech recognition unit in order the generate the phoneme annotation data. In some situations, a transcript of the audio data will be present in the data file. Such an embodiment is illustrated in Figure 14. In this embodiment, the data file 81 represents a digital video file having video data 81-1, audio data 81-2 and script data 81-3 which defines the lines for the various actors in the video film. As shown, the script data 81-3 is passed through a text to phoneme converter 83, which generates phoneme lattice data 85 using a stored dictionary which translates words into possible sequences of phonemes. This phoneme lattice data 85 is then combined with the script data 81-3 to generate the above described phoneme and word lattice annotation data 81-4. This annotation data is then added to the data file 81 to generate an augmented data file 81′ which is then added to the database 29. As those skilled in the art will appreciate, this embodiment facilitates the generation of separate phoneme and word lattice annotation data for the different speakers within the

video data file, since the script data usually contains indications of who is talking. The synchronisation of the phoneme and word lattice annotation data with the video and audio data can then be achieved by performing a forced time alignment of the script data with the audio data using an automatic speech recognition system (not shown).

CLAIMS:

1. An apparatus for generating annotation data for use in annotating a data file comprising audio data, the apparatus comprising:

an automatic speech recognition system for generating phoneme data for the audio data in the data file;

a word decoder for identifying possible words within the phoneme data generated by the automatic speech recognition system; and

means for combining the phoneme data and the decoded words to generate annotation data defining a phoneme and word lattice for the audio data in the data file;

wherein said combining means comprises:

(i) means for generating data defining a plurality of nodes within the lattice and a plurality of links connecting the nodes within the lattice; and

(ii) means for generating data associating a plurality of phonemes of the phoneme data with a respective plurality of links and for associating at least one of the identified words with at least one of said links.

2. An apparatus for generating annotation data for use in annotating a data file comprising text data, the apparatus comprising:

a text to phoneme converter for generating phoneme data for the text data in the data file; and

means for combining the phoneme data and the words in the text data to generate annotation data defining a phoneme and word lattice for the text data in the data file;

wherein said combining means comprises:

(i) means for generating data defining a plurality of nodes within the lattice and a plurality of links connecting the nodes within the lattice; and

(ii) means for generating data associating a plurality of phonemes of the phoneme data with a respective plurality of links and for associating at least one of the identified words with at least one of said links.

3. An apparatus according to claim 1 or 2, wherein said combining means is operable to generate said data defining said phoneme and word lattice in blocks of said nodes.

4. An apparatus according to any preceding claim, wherein said combining means is operable to generate data defining time stamp information for each of said nodes.

5. An apparatus according to claim 4, wherein said combining means is arranged to generate said phoneme and

word lattice data in blocks of equal time duration.

6. An apparatus according to claim 3 or 5, wherein said combining means is operable to generate data which defines each block's location within a database.

7. An apparatus according to claim 4 or any claim dependent thereon, wherein said data file includes a time sequential signal, and wherein said combining means is operable to generate time stamp data which is time synchronised with said time sequential signal.

8. An apparatus according to claim 7, wherein said time sequential signal is an audio and/or video signal.

9. An apparatus according to claim 1 or any claim dependent thereon, wherein said audio data includes audio data which defines the parol of a plurality of speakers, and wherein said combining means is operable to generate data which defines separate phoneme and word lattice annotation data for the parol of each speaker.

10. An apparatus according to claim 2 or any claim dependent thereon, wherein said text data defines the parol of a plurality of speakers, and wherein said combining means is operable to generate data defining separate phoneme and word lattice annotation data for the

parol of each speaker.

11. An apparatus according to claim 1 or any claim dependent thereon, wherein said speech recognition system is operable to generate data defining a weighting for the phonemes associated with said links.

12. An apparatus according to claim 1 or any claim dependent thereon, wherein said word decoder is operable to generate data defining a weighting for the words associated with said links.

13. An apparatus according to any preceding claim, wherein said means for defining a plurality of nodes and a plurality of links is operable to define at least one node which is connected to a plurality of other nodes by a plurality of links.

14. An apparatus according to claim 13, wherein at least one of said plurality of links connecting said node to said plurality of other nodes is associated with a phoneme and wherein at least one of said links connecting said node to said plurality of other nodes is associated with a word.

15. A method of generating annotation data for use in annotating a data file comprising audio data, the method

comprising the steps of:

using an automatic speech recognition system to generate phoneme data for the audio data in the data file;

using a word decoder to identify possible words within the phoneme data generated by the automatic speech recognition system; and

combining the phoneme data and the decoded words to generate annotation data defining a phoneme and word lattice for the audio data in the data file;

wherein said combining step comprises the steps of:

(i) generating data defining a plurality of nodes within the lattice and a plurality of links connecting the nodes within the lattice; and

(ii) generating data associating a plurality of phonemes of the phoneme data with a respective plurality of links and for associating at least one of the identified words with at least one of said links.

16. A method of generating annotation data for use in annotating a data file comprising text data, the method comprising the steps of:

using a text to phoneme converter to generate phoneme data for the text data in the data file; and

combining the phoneme data and the words in the text data to generate annotation data defining a phoneme and word lattice for the text data in the data file;

wherein said combining step comprises the steps of:

(i) generating data defining a plurality of nodes within the lattice and a plurality of links connecting the nodes within the lattice; and

(ii) generating data associating a plurality of phonemes of the phoneme data with a respective plurality of links and for associating at least one of the identified words with at least one of said links.

17. A method according to claim 15 or 16, wherein said combining step generates said data defining said phoneme and word lattice in blocks of said nodes.

18. A method according to any of claims 15 to 17, wherein said combining step generates data defining time stamp information for each of said nodes.

19. A method according to claim 18, wherein said combining step generates said phoneme and word lattice data in blocks of equal time duration.

20. A method according to claim 17 or 19, wherein said combining step generates data which defines each block's location within a database.

21. A method according to claim 18 or any claim dependent thereon, wherein said data file includes a time

sequential signal, and wherein said combining step generates time stamp data which is time synchronised with said time sequential signal.

22. A method according to claim 21, wherein said time sequential signal is an audio and/or video signal.

23. A method according to claim 15 or any claim dependent thereon, wherein said audio data includes audio data which defines the parol of a plurality of speakers, and wherein said combining step generates data which defines separate phoneme and word lattice annotation data for the parol of each speaker.

24. A method according to claim 16 or any claim dependent thereon, wherein said text data defines the parol of a plurality of speakers, and wherein said combining step generates data defining separate phoneme and word lattice annotation data for the parol of each speaker.

25. A method according to claim 15 or any claim dependent thereon, wherein said speech recognition system generates data defining a weighting for the phonemes associated with said links.

26. A method according to claim 15 or any claim

dependent thereon, wherein said word decoder generates data defining a weighting for the words associated with said links.

27. A method according to any of claims 15 to 26, wherein said step of defining a plurality of nodes and a plurality of links defines at least one node which is connected to a plurality of other nodes by a plurality of links.

28. A method according to claim 27, wherein at least one of said plurality of links connecting said node to said plurality of other nodes is associated with a phoneme and wherein at least one of said links connecting said node to said plurality of other nodes is associated with a word.

ABSTRACT

DATABASE ANNOTATION AND RETRIEVAL

A data structure is provided for annotating data files within a database. The annotation data comprises a phoneme and word lattice which allows the quick and efficient searching of data files within the database, in response to a user's input query for desired information. The structure of the annotation data is such that it allows the input query to be made by voice and can be used for annotating various kinds of data files, such as audio data files, audio and visual data files, multimedia data files etc.

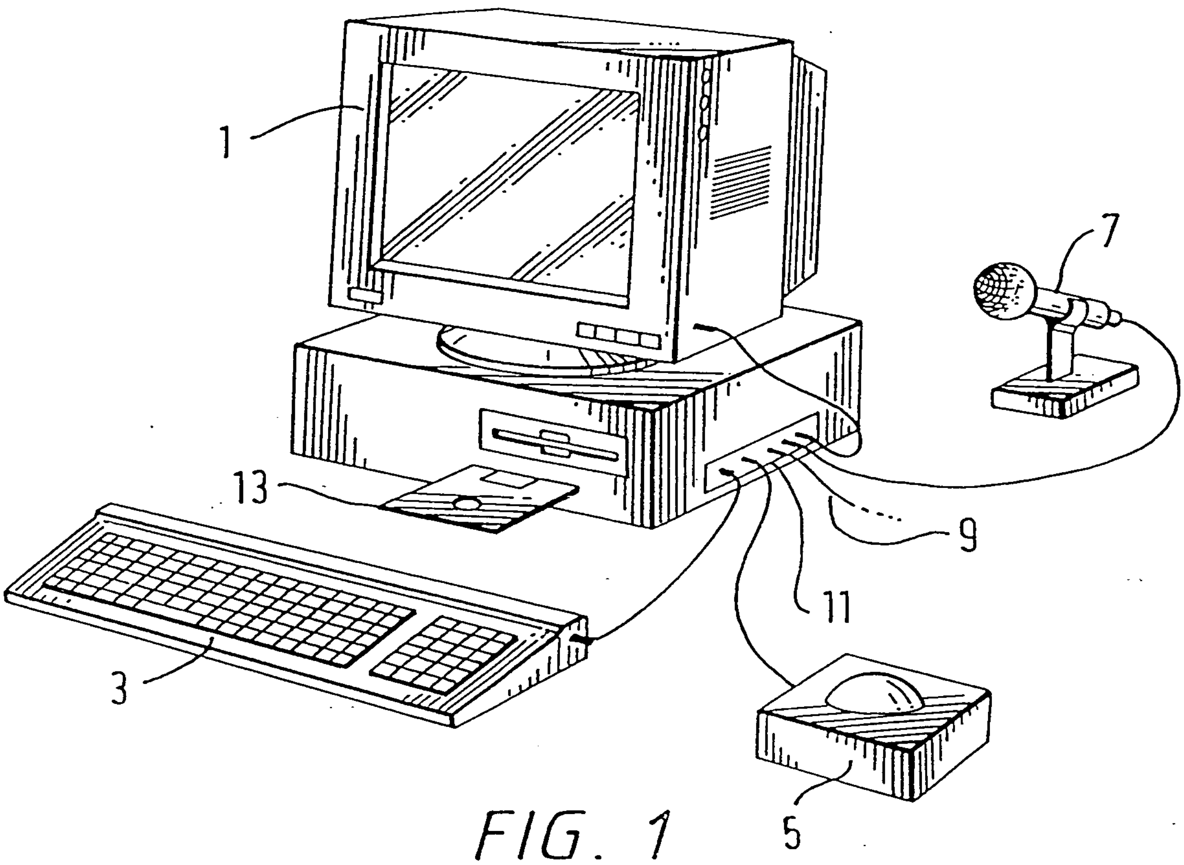


FIG. 1

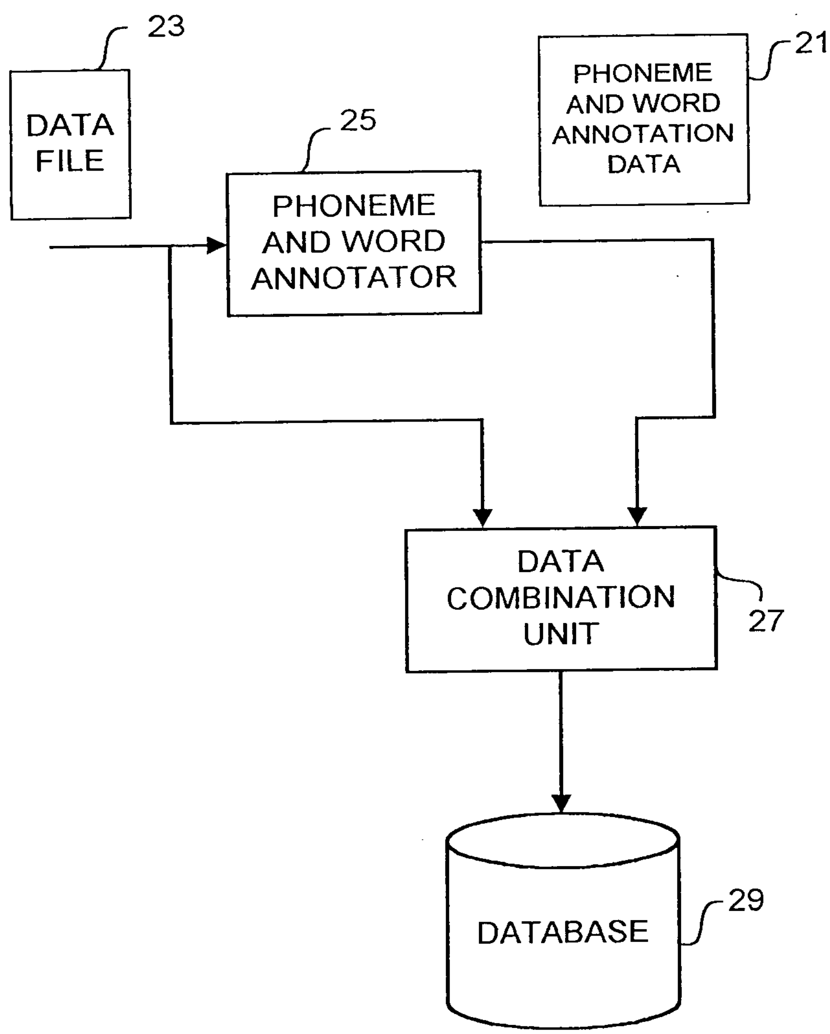
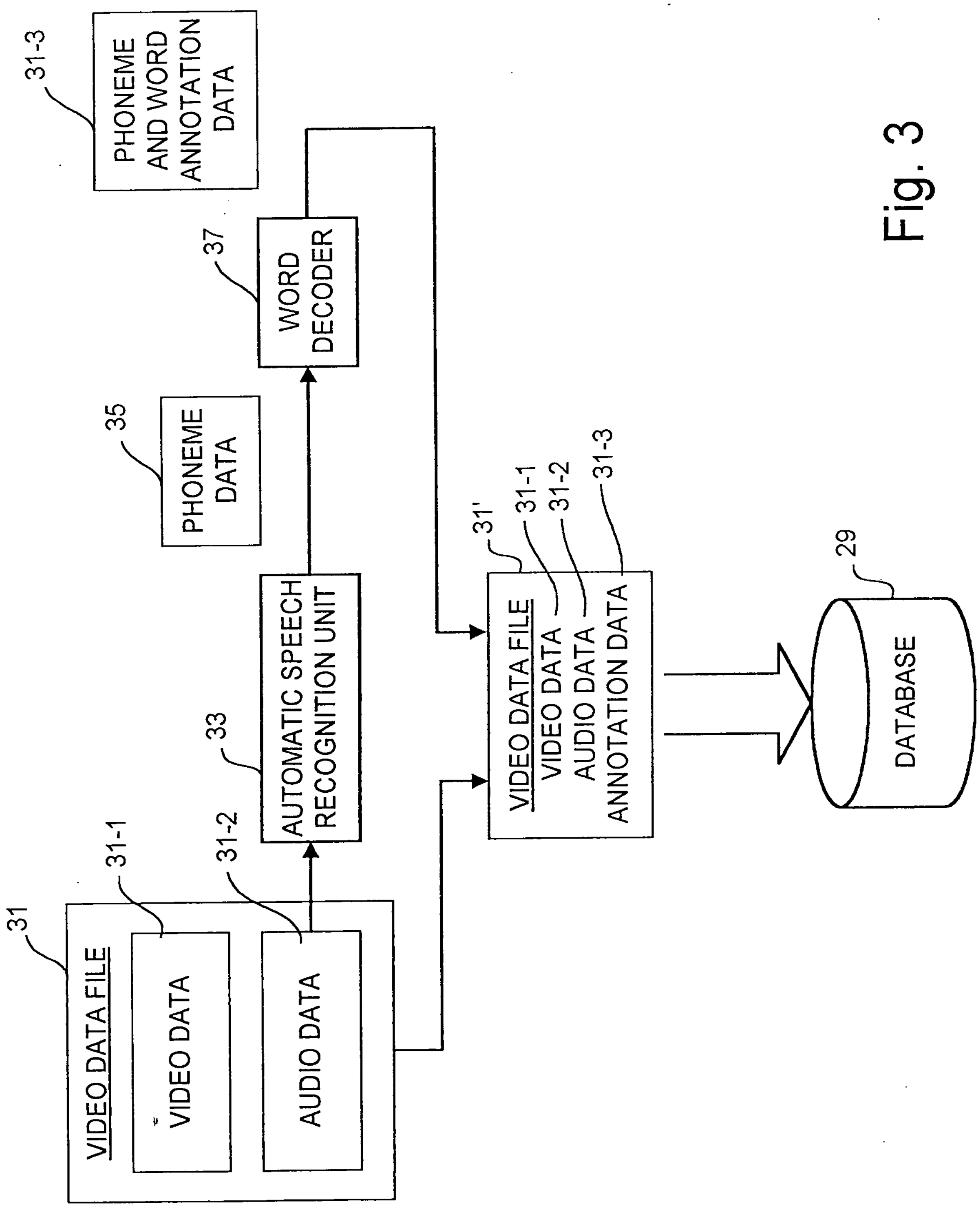

23
DATA FILE
25
PHONEME AND WORD ANNOTATOR
21
PHONEME AND WORD ANNOTATION DATA
DATA COMBINATION UNIT
27
DATABASE
29

Fig. 2



31
VIDEO DATA FILE
31-1
VIDEO DATA
31-2
AUDIO DATA
33
AUTOMATIC SPEECH RECOGNITION UNIT
35
PHONEME DATA
37
WORD DECODER
31-3
PHONEME AND WORD ANNOTATION DATA
31'
31-1
31-2
31-3
VIDEO DATA FILE
VIDEO DATA
AUDIO DATA
ANNOTATION DATA
29
DATABASE

Fig. 3

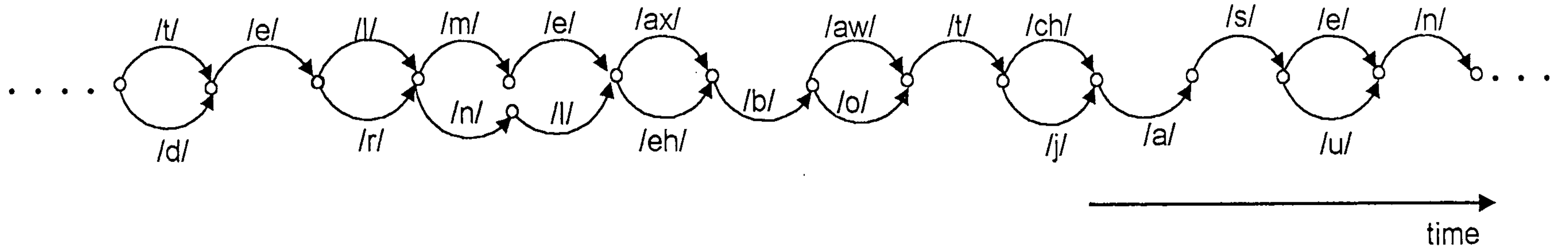

/t/
/d/
/e/
/l/
/r/
/m/
/n/
/e/
/l/
/ax/
/eh/
/b/
/aw/
/o/
/t/
/ch/
/j/
/a/
/s/
/e/
/u/
/n/
time

Fig. 4a

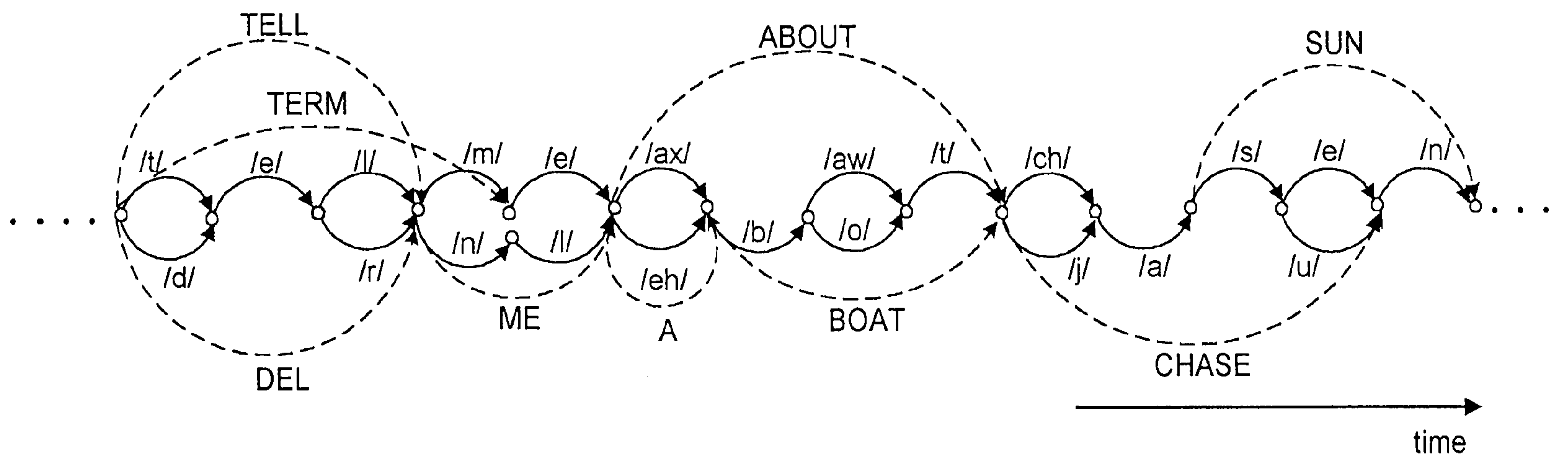

TELL
TERM
DEL
/t/
/d/
/e/
/l/
/r/
/m/
/n/
/e/
/l/
ME
/ax/
/eh/
A
ABOUT
/b/
/aw/
/o/
/t/
BOAT
/ch/
/j/
/a/
CHASE
SUN
/s/
/e/
/u/
/n/
time

Fig. 4b

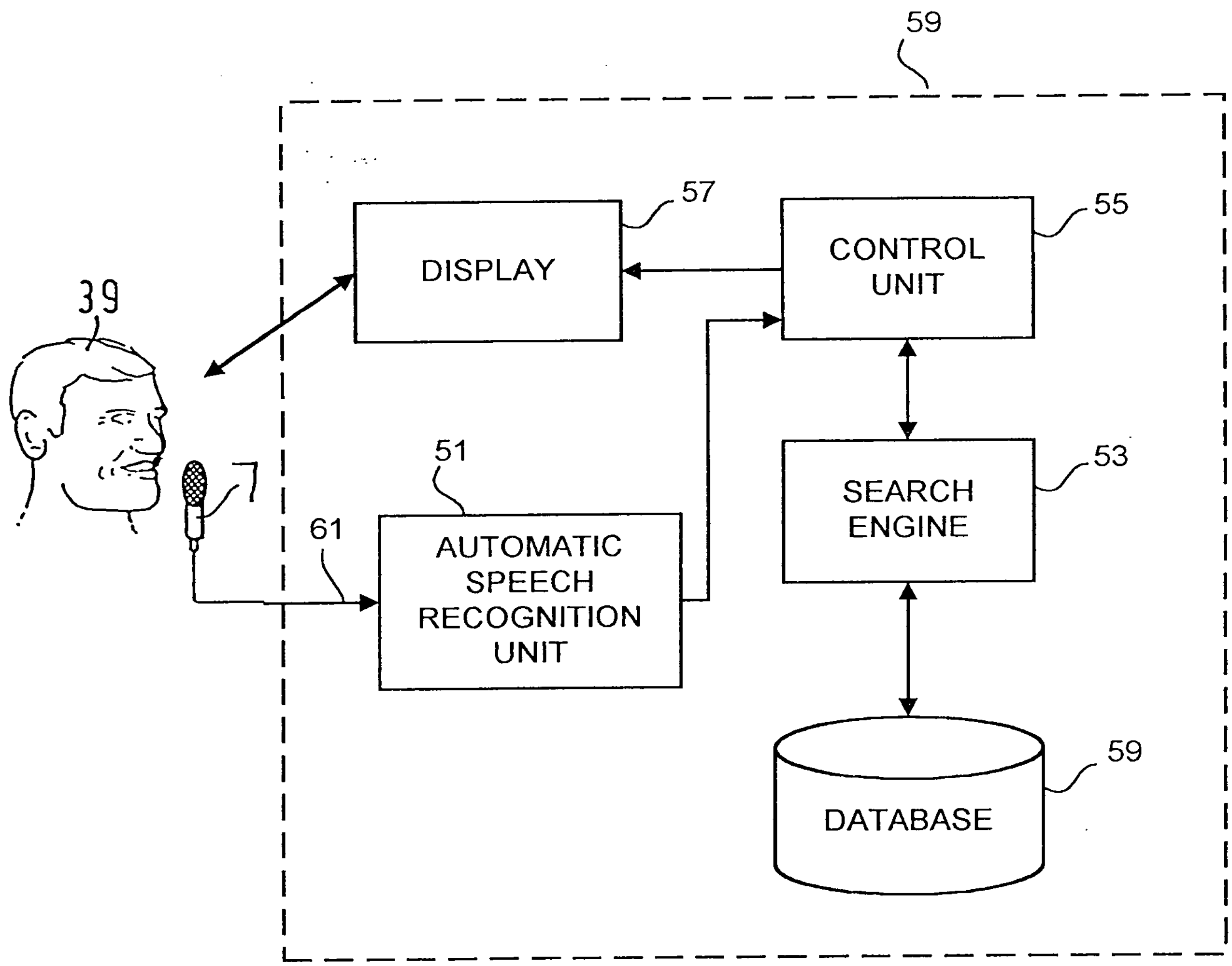

59
57
DISPLAY
CONTROL UNIT
55
39
SEARCH ENGINE
53
51
7
61
AUTOMATIC SPEECH RECOGNITION UNIT
DATABASE
59

Fig. 5

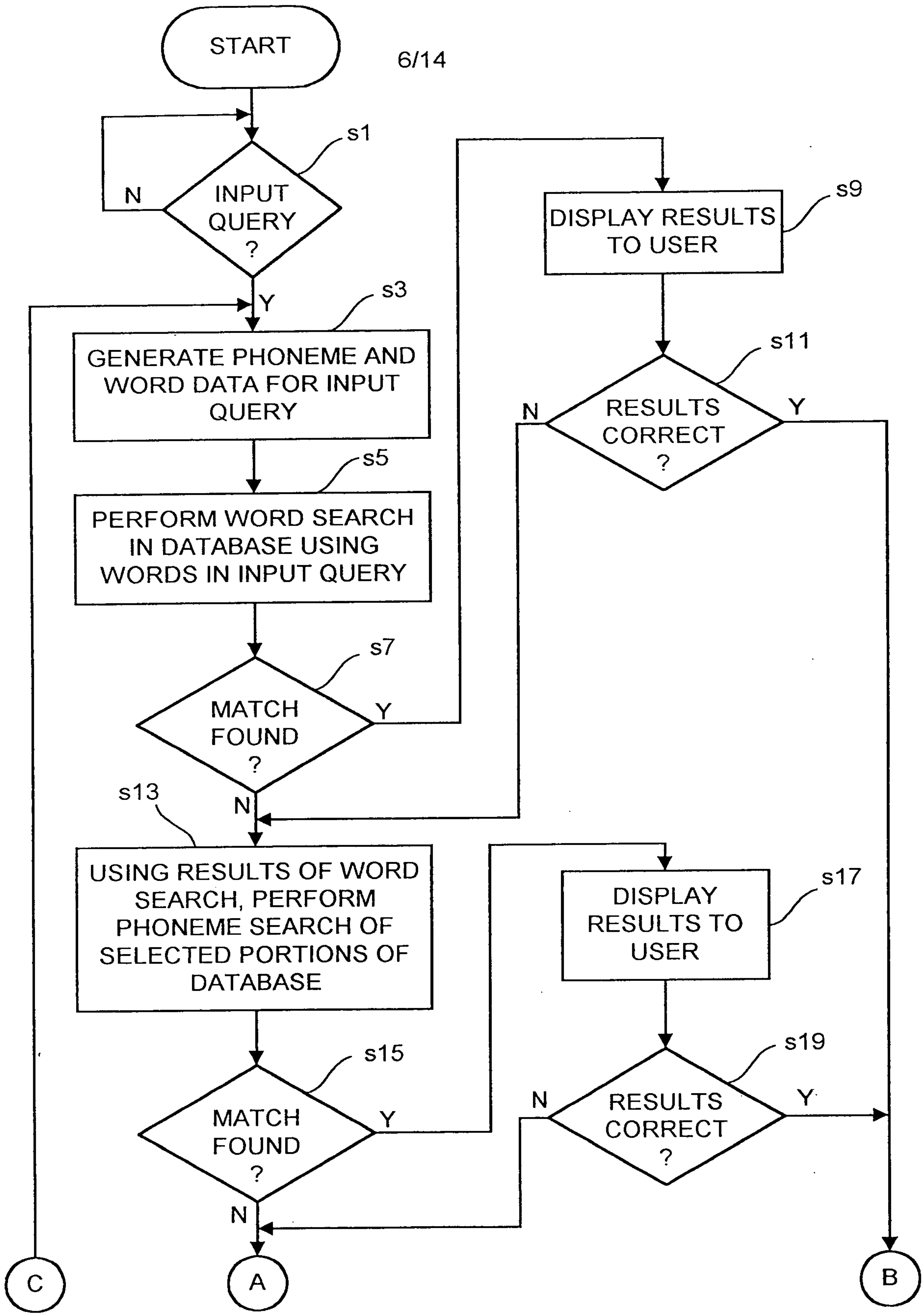

START
INPUT QUERY ?
s1
N
Y
GENERATE PHONEME AND WORD DATA FOR INPUT QUERY
s3
PERFORM WORD SEARCH IN DATABASE USING WORDS IN INPUT QUERY
s5
MATCH FOUND ?
s7
Y
N
DISPLAY RESULTS TO USER
s9
RESULTS CORRECT ?
s11
N
Y
USING RESULTS OF WORD SEARCH, PERFORM PHONEME SEARCH OF SELECTED PORTIONS OF DATABASE
s13
MATCH FOUND ?
s15
Y
N
DISPLAY RESULTS TO USER
s17
RESULTS CORRECT ?
s19
N
Y
C
A
B

Fig. 6a

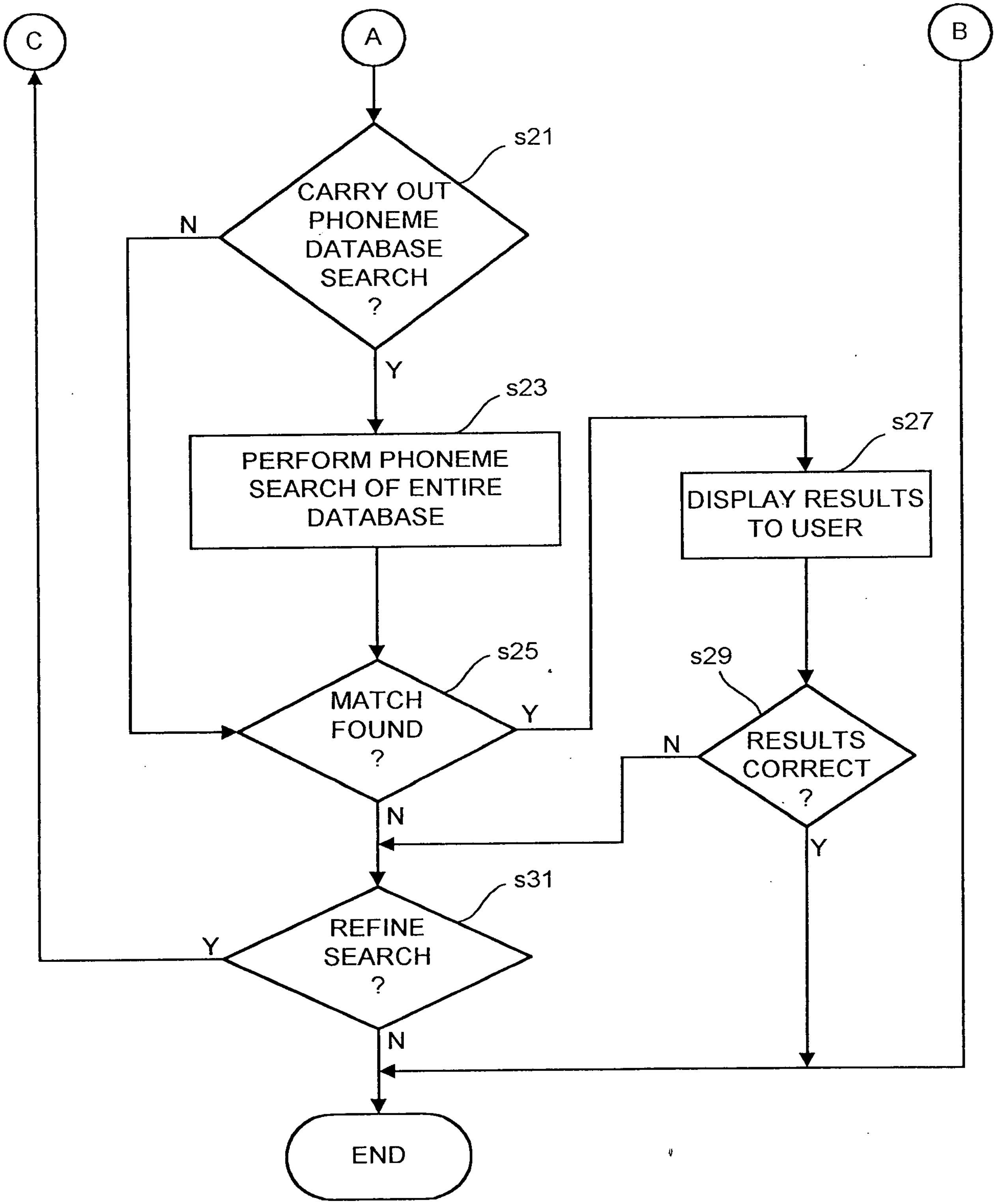

C
A
B
s21
CARRY OUT PHONEME DATABASE SEARCH ?
N
Y
s23
PERFORM PHONEME SEARCH OF ENTIRE DATABASE
s27
DISPLAY RESULTS TO USER
s25
MATCH FOUND ?
Y
N
s29
RESULTS CORRECT ?
N
Y
s31
REFINE SEARCH ?
Y
N
END

Fig. 6b

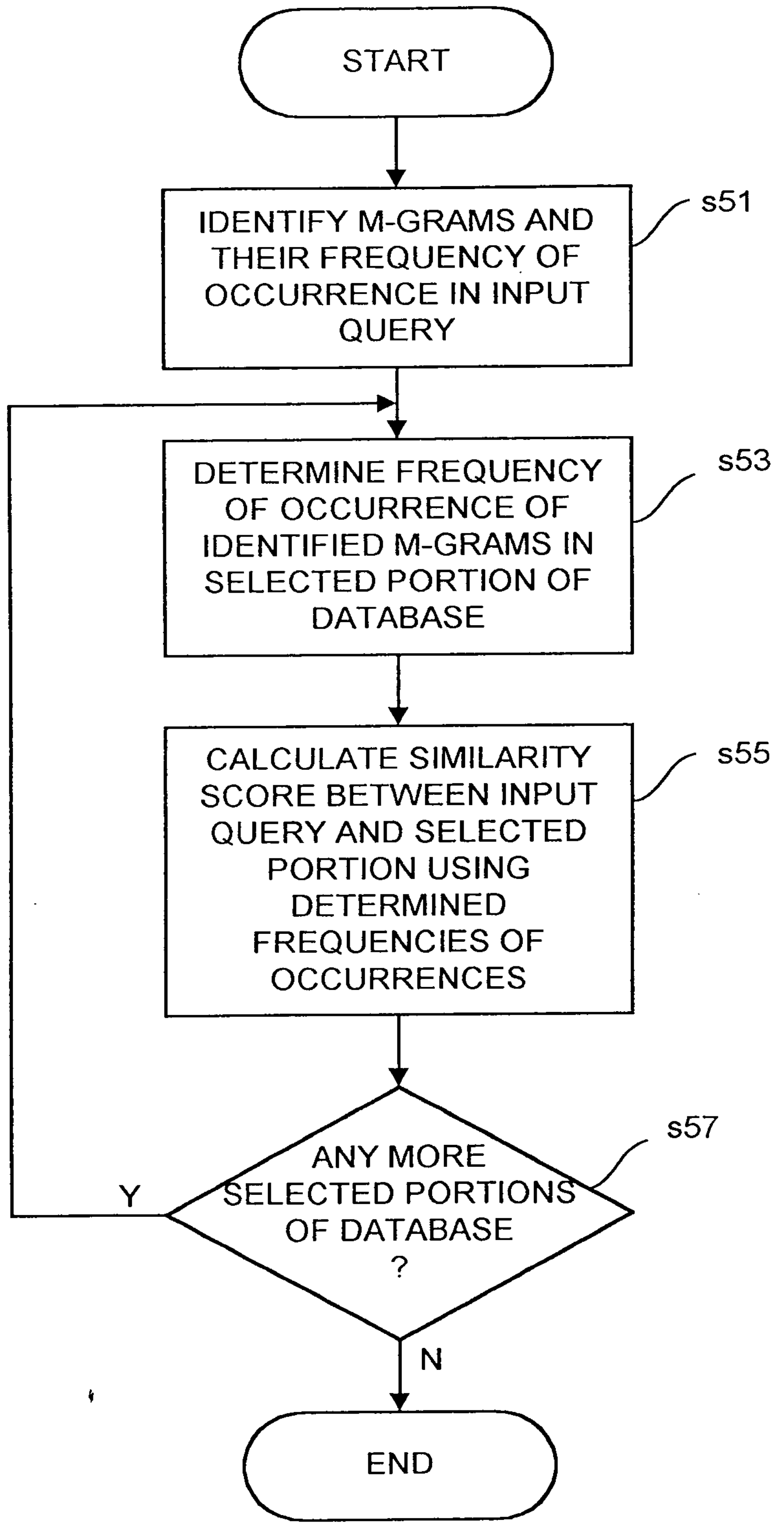

START
IDENTIFY M-GRAMS AND THEIR FREQUENCY OF OCCURRENCE IN INPUT QUERY
s51
DETERMINE FREQUENCY OF OCCURRENCE OF IDENTIFIED M-GRAMS IN SELECTED PORTION OF DATABASE
s53
CALCULATE SIMILARITY SCORE BETWEEN INPUT QUERY AND SELECTED PORTION USING DETERMINED FREQUENCIES OF OCCURRENCES
s55
ANY MORE SELECTED PORTIONS OF DATABASE ?
s57
Y
N
END

Fig.7

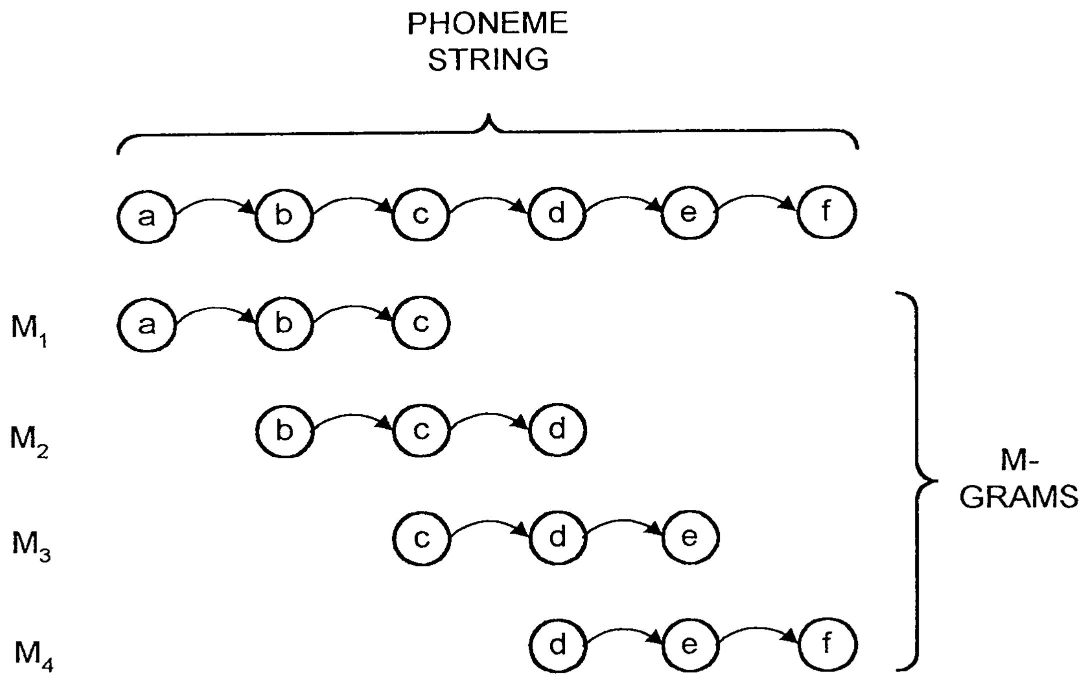


Fig. 8

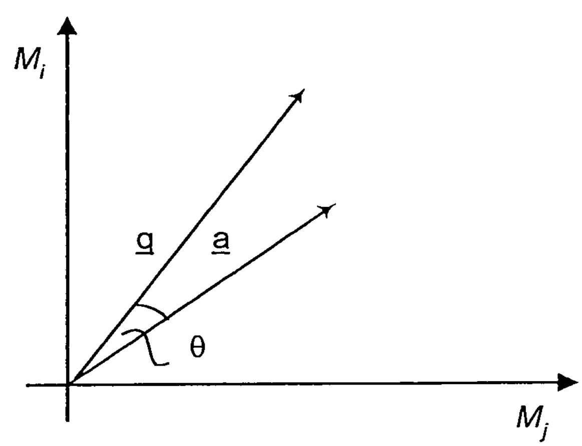


Fig. 9

THIS PAGE BLANK (USPTO)

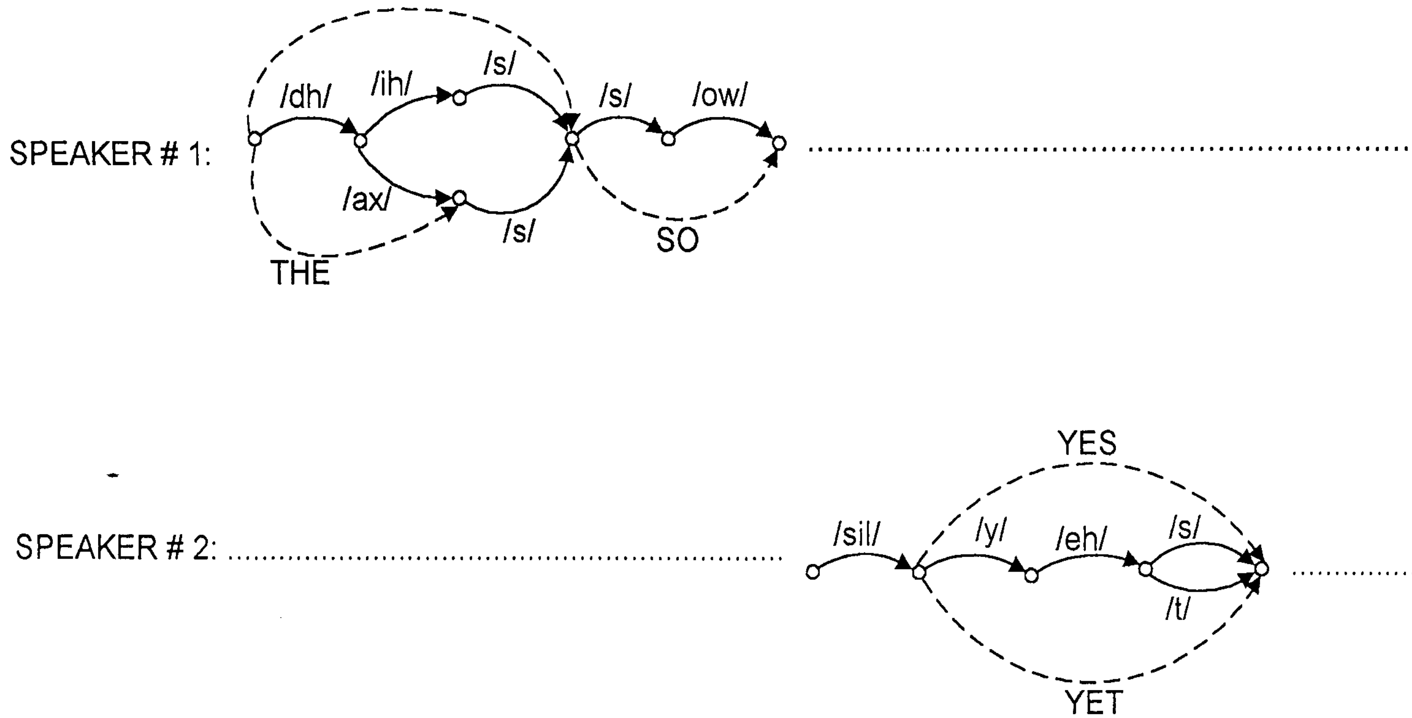


Fig. 10

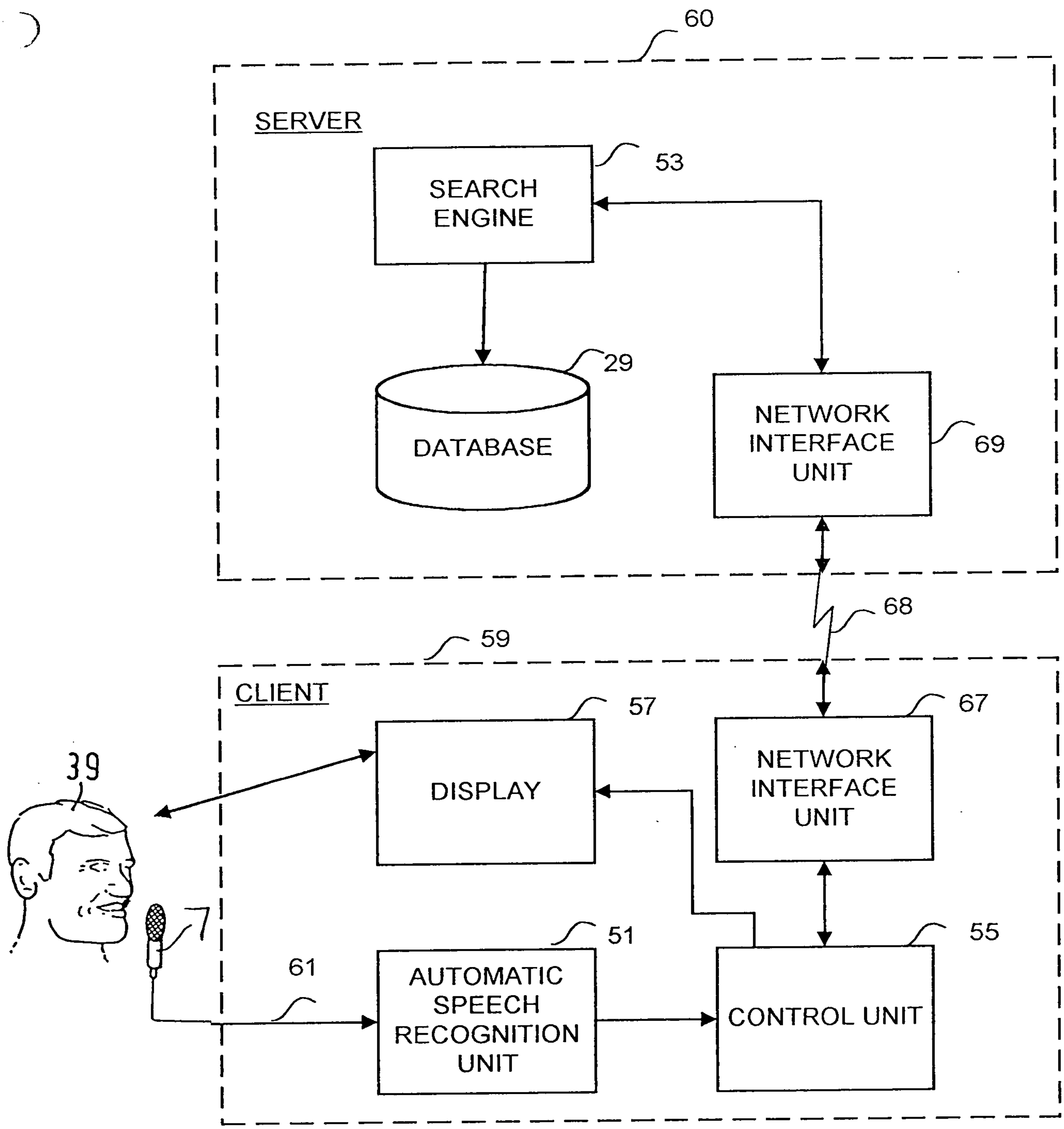

60
SERVER
SEARCH ENGINE
53
29
DATABASE
NETWORK INTERFACE UNIT
69
68
59
CLIENT
57
DISPLAY
67
NETWORK INTERFACE UNIT
39
7
51
55
61
AUTOMATIC SPEECH RECOGNITION UNIT
CONTROL UNIT

Fig. 11

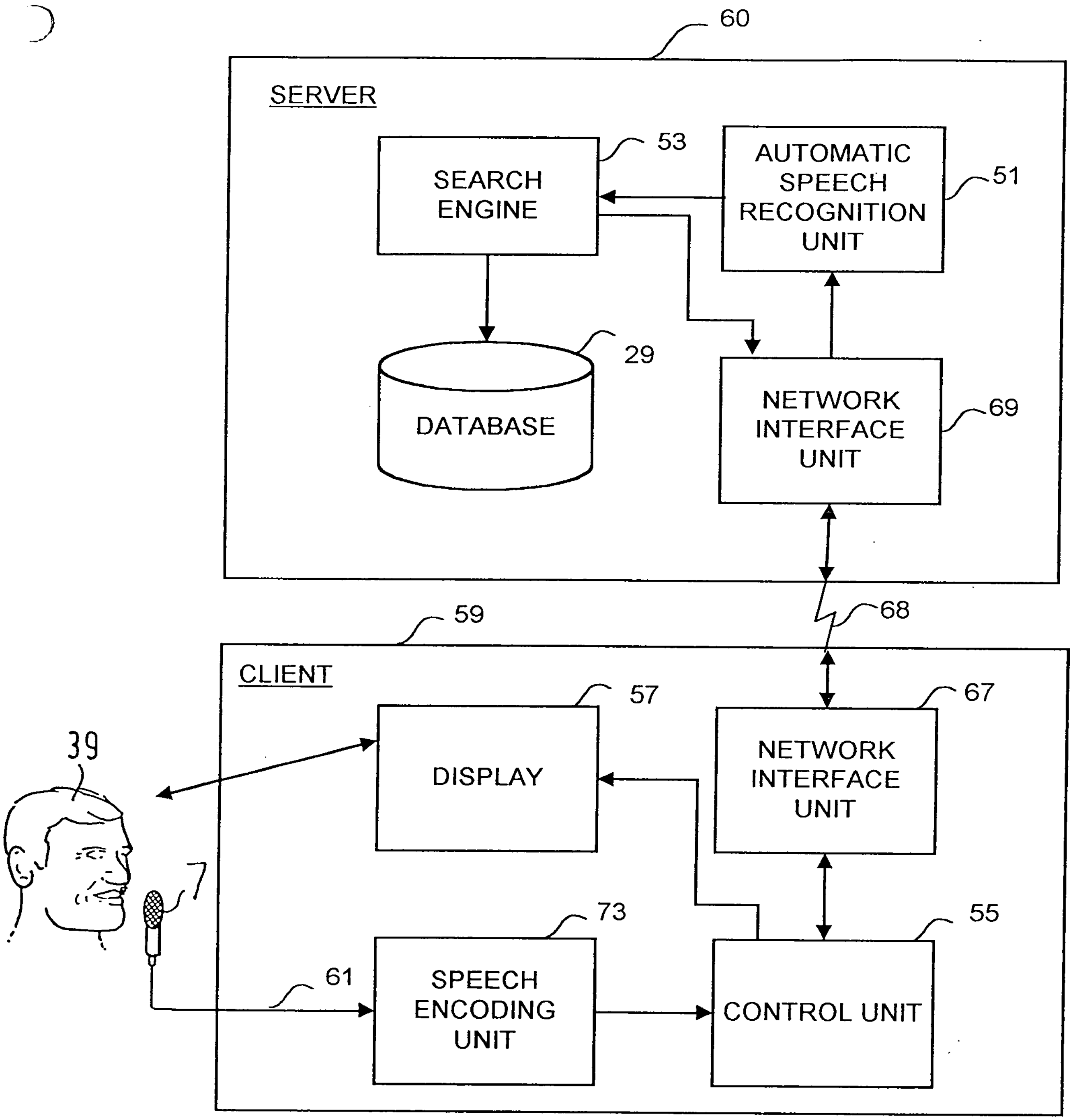

60
SERVER
SEARCH ENGINE
53
AUTOMATIC SPEECH RECOGNITION UNIT
51
DATABASE
29
NETWORK INTERFACE UNIT
69
68
59
CLIENT
57
DISPLAY
NETWORK INTERFACE UNIT
67
39
7
73
55
SPEECH ENCODING UNIT
CONTROL UNIT
61

Fig. 12

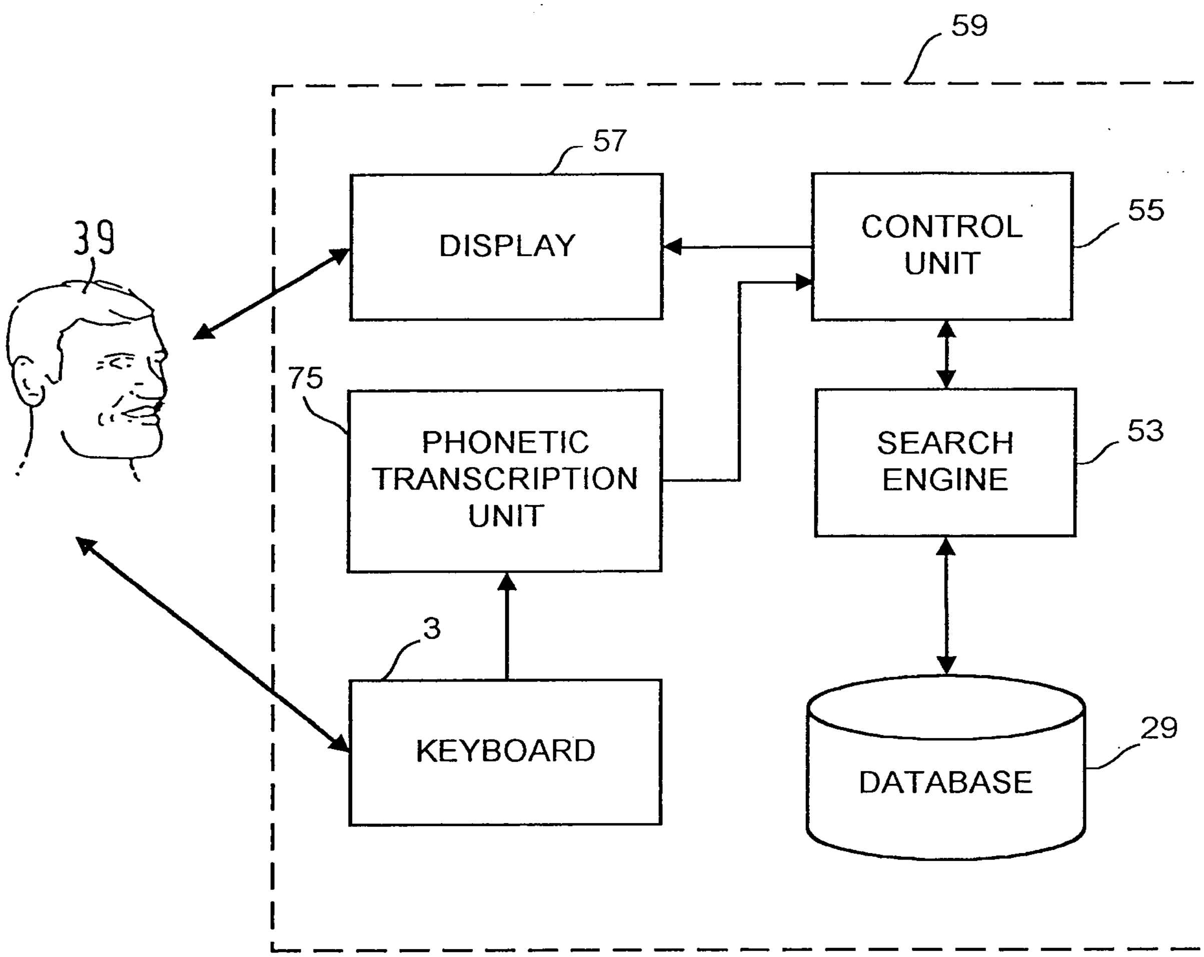

59
57
DISPLAY
CONTROL UNIT
55
39
75
PHONETIC TRANSCRIPTION UNIT
SEARCH ENGINE
53
3
KEYBOARD
DATABASE
29

Fig. 13

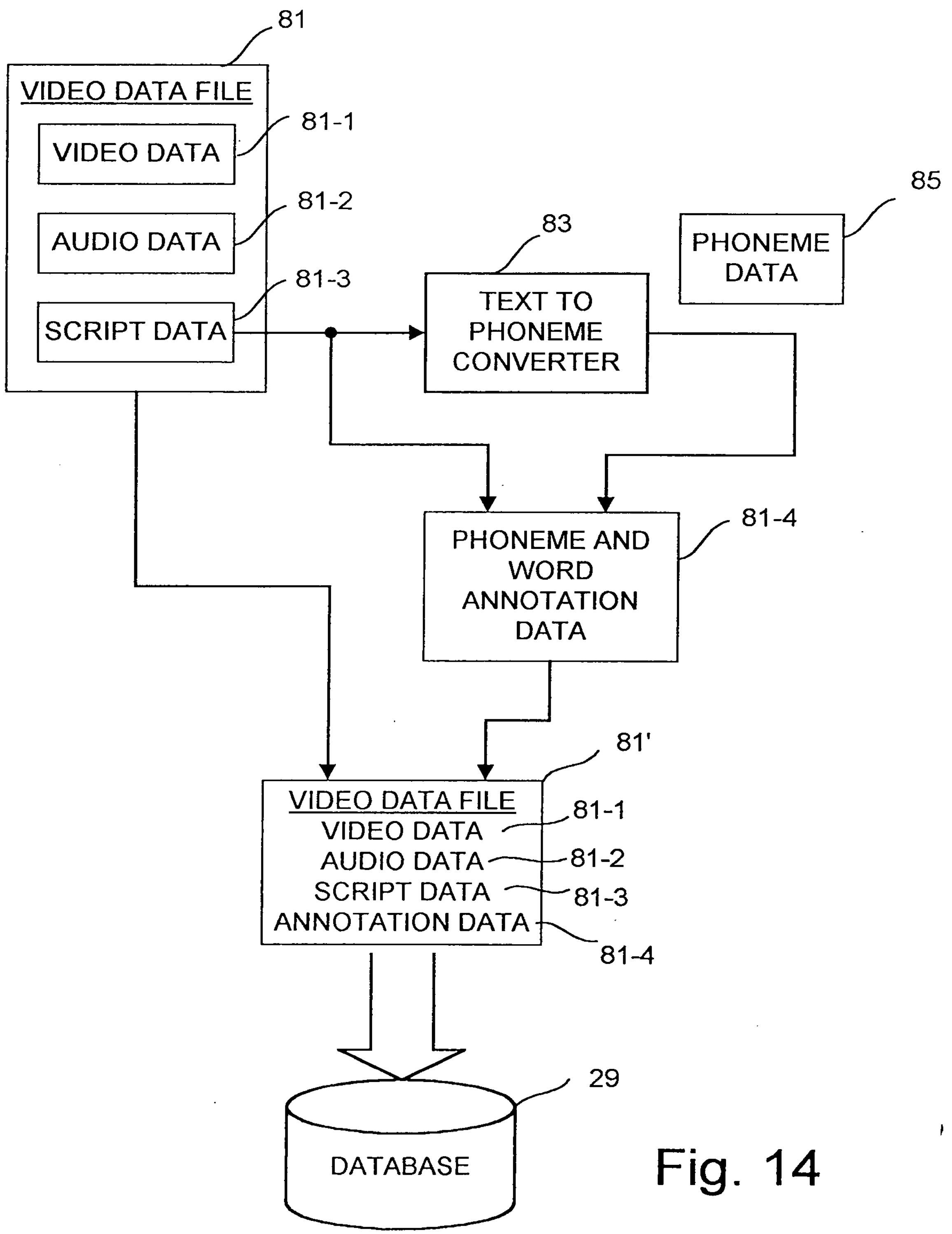

81
VIDEO DATA FILE
VIDEO DATA
81-1
AUDIO DATA
81-2
SCRIPT DATA
81-3
83
TEXT TO PHONEME CONVERTER
PHONEME DATA
85
PHONEME AND WORD ANNOTATION DATA
81-4
81'
VIDEO DATA FILE
VIDEO DATA
81-1
AUDIO DATA
81-2
SCRIPT DATA
81-3
ANNOTATION DATA
81-4
29
DATABASE

Fig. 14